(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) Publication number : **0 510 632 A2**

(12) EUROPEAN PATENT APPLICATION

(21) Application number : **92106933.2**

(51) Int. Cl.^{5} : **G10L 5/06**

(22) Date of filing : **23.04.92**

(30) Priority : **24.04.91 JP 119086/91**

(43) Date of publication of application :
28.10.92 Bulletin 92/44

(84) Designated Contracting States :
DE FR GB

(71) Applicant : **NEC CORPORATION**
7-1, Shiba 5-chome Minato-ku
Tokyo 108-01 (JP)

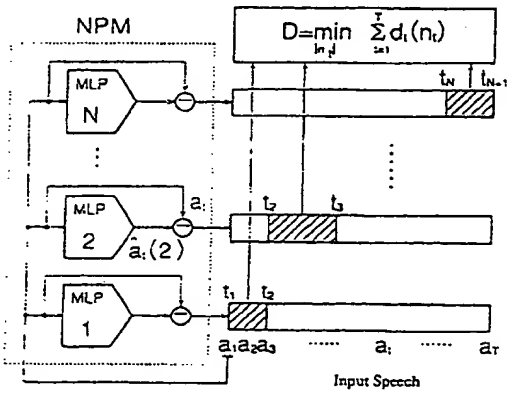
(72) Inventor : **Iso, Ken-ichi, c/o NEC Corporation**
7-1, Shiba 5-chome
Minato-ku, Tokyo (JP)

(74) Representative : **Vossius & Partner**
Siebertstrasse 4 P.O. Box 86 07 67
W-8000 München 86 (DE)

(54) Speech recognition by neural network adapted to reference pattern learning.

(57) A speech recognition method according to the present invention uses distances calculated through a variance weighting process using covariance matrixes as the local distances (prediction residuals) between the feature vectors of input syllables/sound elements and predicted vectors formed by different statuses of reference neural prediction models (NPM's) using finite status transition networks. The category to minimize the accumulated value of these local distances along the status transitions of all the prediction models is figured out by dynamic programming, and used as the recognition output. Learning of the reference prediction models used in this recognition method is accomplished by repeating said distance calculating process and the process to correct the parameters of the different statuses and the covariance matrixes of said prediction models in the direction of reducing the distance between the learning patterns whose category is known and the prediction models of the same category as this known category, and what have satisfied prescribed conditions of convergence through these calculating and correcting processes are determined as reference pattern models.

Fig. 3



EP 0 510 632 A2

The present invention relates to a speech recognition method, and more particularly to a speech recognition method manifesting a high rate of recognition without requiring learning with a particularly large quantity of training data.

Speech signals are expressed in time series patterns of feature vectors, and speech recognition is based on the degree of identity between a reference pattern representing a known piece of speech and the pattern of the input speech signal. For these time series patterns, the Hidden Markov Model (HMM) is extensively used as described in detail in the specifications of the U.S. Patents Nos. 4,587,670 and 4,582,180. The HMM itself will not be explained in detail here because its detailed description can be found in S.E. Levinson, "Structural Method in Automatic Speech Recognition", Proc. IEEE, 73, No. 11, 1985, pp. 1625-1650, besides said U.S. Patents.

The HMM assumes that the time series of feature vectors are generated by the Markov probability process. The standard patterns of the HMM are represented in a plurality of statuses and transitions between the statuses, and each status outputs a feature vector according to a predetermined distribution of probability density while each transition between statuses is accompanied by a predetermined probability of transition. The likelihood, which represents the degree of matching between the input pattern and a reference pattern, is given by the probability of the Markov probability model to generate a series of input pattern vectors. The probability of transition between statuses and the parameter to define the function of probability density distribution, which characterize each reference pattern, can be determined with the Baum Welch algorithm using a plurality of sets of vocalization data for the training purpose.

However, the Baum Welch algorithm, which is a statistical learning method, requires a large quantity of training data to determine the parameters of the model corresponding to reference patterns. Therefore, the load of vocalization is extremely great when a speech recognition apparatus begins to be newly used, and this presents a serious obstacle to the practical use of such apparatuses. Therefore, with a view to reducing this load, a number of speaker-adaptive methods have already been proposed to adapt a speech recognition apparatus to the speaker with a relatively small quantity of training data.

A speaker-adaptive method defines the similarity of acoustic events according to reference patterns corresponding to known speech signals and a new speaker's vocalization data for adaptation, basically using the physical distance between feature vectors as the scale, and carries out adaptation by estimating, on the basis of that similarity, the parameters of the model corresponding to acoustic events absent in the vocalization data for adaptation.

However, such a method of adaptation based on an estimation relying solely on physical distances, though providing a somewhat higher rate of recognition than before the adaptation, is far less effective in recognition than a method using reference patterns corresponding to a specific speaker, consisting of a large quantity of speech data. (For further details, see K. Shikano, K.F. Lee and R. Reddy, "Speaker Adaptation through Vector Quantization", Proc. ICASSP-86, Tokyo, 1986, pp. 2643-2646.)

Meanwhile, as means for improving the rate of recognition, the inventors of the present invention proposed a pattern recognition method based on the prediction of the aforementioned time series patterns. Using multilayer perceptrons (MPL's) based on a neural network as predictive means for the time series patterns, the outputs of the MLP's constitute reference patterns. The inventors named the reference patterns the "neural prediction model" (NPM). This NPM will not be described in detail here as its detailed explanations can be found in K. Iso and T. Watanabe, "Speaker-Independent Word Recognition Using a Neural Prediction Model," Proc. ICASSP-90, New Mexico, 1990, p. 441-444 and the pending U.S. Patent Ser. No. (07-521625). In the NPM described in these references, a predictor (MLP) in the nth status of a reference pattern model consisting of a finite status transition network calculates a predicted for the feature vector of the input patterns at time t from a plurality of feature vectors at time $t-1$ and before. The distance between this predicted vector and the feature vector of the input pattern at time t is supposed to be the local distance between said two feature vectors. In the NPM described in the above cited references, the squared distance or the like between the vectors is used as this local distance.

BRIEF SUMMARY OF THE INVENTION

Object of the Invention

An object of the present invention is to reduce, in relative terms, the contributions of components inferior in predictive accuracy (i.e. more fluctuating) out of the components of said predicted vectors, thereby increase the predictive accuracy of the predictor and improve the accuracy of recognition by the NPM.

Summary of the Invention

A pattern recognition method according to the invention recognizes the time series patterns of feature vectors representing input speech signals by using the NPM constituting said finite status transition network. Each status of this finite status transition network has a predictor for calculating a predicted vector from a plurality of feature vectors of the input time series patterns at time t-1 and before and a plurality of feature vectors at time t+1 and after. This predicted vector is compared with the feature vector of the input time series patterns at time t. As said local distance indicated by the result of this comparison, i.e. the local distance between the feature vector of the input time series patterns at time t (input feature vector) and the nth status of the finite status transition network (predicted feature vector), there is used the prediction residual calculated from the input feature vector, the predicted feature vector and a covariance matrix accom panying said nth status in advance. The total difference between said input time series patterns and the reference pattern model is given by a cumulative value following said status transition of said local distance. This cumulative value is calculated for every category of the reference pattern model, and the category having the smallest cumulative value is selected as the recognition output.

The NPM according to the present invention composes said reference pattern model by learning. First, the initial values are set for the parameters of said predictor and covariance matrix accompanying each status of said finite status transition network. Next, said total distance between the learning pattern whose category is known and the reference pattern model of the same category as said category is calculated, and the parameters of the predictor and covariance matrix of each state are corrected in the direction of reducing said total distance without fail by a predetermined algorithm. This correction is repeated, and the pattern model satisfying predetermined conditions of convergence is eventually selected as the reference pattern model.

Brief Description of the Drawings

The above-mentioned and other objects, features and advantages of the present invention will become more apparent by reference to the following detailed description of the invention taken in conjunction with the accompanying drawings, wherein:

FIG. 1 illustrates the configuration of the multilayer perceptrons (MLP's) used as the predictor in the invention;

FIG. 2 illustrates the finite status transitions of an NPM, which constitutes the reference pattern model according to the invention;

FIG. 3 illustrates the configuration of the recognition algorithm according to the invention;

FIG. 4 is a recognition flow chart illustrating the pattern recognition method according to the invention;

FIG. 5 is a detailed flow chart of the initializing section of FIG. 4;

FIG. 6 is a detailed flow chart of the local distance calculation in FIG. 4;

FIG. 7 is a flow chart illustrating the reference pattern learning method according to the invention;

FIG. 8 is a block diagram of a speech recognition apparatus which is a preferred embodiment of the invention;

FIG. 9 is a detailed flow chart of predicted vector calculation at step 601 in FIG. 6;

FIG. 10 is a detailed flow chart of local distance calculation at step 602 in FIG. 6;

FIG. 11 is a detailed flow chart of initializing at step 701 in FIG. 7;

FIG. 12 is a detailed flow chart of optimal trajectory calculation at step 704 in FIG. 7;

FIG. 13 is a detailed flow chart of the calculation of the quantities of parameter correction at step 706 in FIG. 7;

FIG. 14 is a detailed flow chart of covariance matrix calculation at step 711 in FIG. 7; and

FIG. 15 is a detailed flow chart of convergence decision at step 712 in FIG. 7.

GENERAL DESCRIPTION

To explain the basic principle of speech recognition according to the present invention with reference to FIG. 1, said predictor used in the invention consists of MLP's. As described in detail in M. Funahashi, "On the Approximate Realization of Continuous Mappings by Neural Networks", Neural Networks, Vol. 2, 1989, pp. 183-192, MLP's approximate any (nonlinear) continuous function at any desired accuracy.

In the figure, the time series patterns to be inputted to the MPL's consist feature vectors $a_{t-\gamma F}$, ..., a_{t-1} for "forward prediction" and a_{t+1}, ..., $a_{t+\gamma B}$ for "backward prediction". The latter's prediction backward on the time axis is added to the former's forward prediction to improve the predictive accuracy of the time series patterns which have close correlation backward on the time axis. As the plosive part of a plosive sound, for instance,

is more closely correlated to the transitional part to the following vowel than to the closed section before the plosion, this backward prediction proves effective for plosive sounds.

The output pattern of MLP's is the predicted vector $\hat{a}_t$ for the feature vector a_t of input speech at time t. This predicted vector can be represented by the following equations, using the input-output relationship of MLP's:

$$h_t = f\left(\sum_{s=1}^{\gamma_F} W_s^F a_{t-s} + \sum_{s=1}^{\gamma_B} W_s^B a_{t+s} + \Theta_1\right) \quad (1)$$

$$\hat{a}_t = W_0 h_t + \Theta_0 \quad (2)$$

where W_0, W_1^F ..., $W_{\gamma F}^F$, W_1^B, ..., $W_{\gamma B}^B$ are the matrix of coupling coefficients between the MLP units; Θ_0 and Θ_1, threshold vectors; and f(·), a vector obtained by applying the sigmoid function to each component of the subtrahend vector.

As stated above, composing a predictor of MLP's makes it possible to describe the causal relationship between proximate features vectors in the time series of speech feature vectors as a nonlinear image formed by the MLP's. The relative accuracy of the prediction can be evaluated by using the predicted vector $\hat{a}_t$, which is the output of the MLP's, and the feature vector a_t of the actual input speech as the prediction residual.

An NPM which constitutes the reference pattern model of a basic unit of speech recognition, such as the word or the syllable, is represented by a transition network of finite (in this case four) statuses (finite statuses including 201 through 204), and each status consists of said predictor composed of MLP's. An NPM representing a greater unit (such as the sentence) can be composed by connecting many NPM's for basic recognition units.

Next, the recognition algorithm using the NPM's basically derives from pattern matching between input speech and a reference pattern model. The reference pattern model for discrete recognition is an NPM for a basic recognition unit, while that for continuous recognition is an NPM obtained by connecting basic unit NPMs, and in both cases it is a finite status transition network accompanied by said MLP predictor. According to the present invention, continuous recognition is accomplished by taking note of the sound elements of speech signals, and therefore of the close correlation on the time axis of said feature vectors. For this reason, said finite status transition network is composed in a left-to-right pattern as shown in FIG. 2.

The distance (local distance) $d_t(n)$ between the feature vector a_t of input speech at time t and the nth status of an NPM is given by the following equation:

$$d_t(n) = (a_t - \hat{a}_t(n))^T \Sigma_n^{-1} (a_t - \hat{a}_t(n)) + \ln|\Sigma_n| \quad (3)$$

where $\hat{a}_t(n)$ is the predicted vector by the MLP predictor in the nth status and

$$\Sigma_n,$$

the covariance matrix in the nth status. The prediction residual is represented by $d_t(n)$, and

$$\Sigma_n$$

in the equation is a quantity introduced to normalize the different extents of fluctuation of the prediction residual from component to component of the feature vector. Equation (3) can be interpreted as the logarithmic probability obtained when the probability at which the feature vector a_t is observed in the nth status of the NPM is approximated by a Gaussian distribution represented by the following equation:

$$(P(n|a_t) \sim \frac{1}{\sqrt{(2\pi)^c |\Sigma_n|}} \exp\left[-\frac{1}{2}(a_t - \hat{a}_t(n))^T \Sigma_n^{-1} (a_t - \hat{a}_t(n))\right] \quad (4)$$

If the nondiagonal terms of the covariance matrix

$$\Sigma_n$$

are negligibly small, equation (3) can be approximated by the following equation:

$$d_t(n) = \sum_{c=1}^{C} \frac{(a_{tc} - \hat{a}_{tc}(n))^2}{\sigma_{nc}^2} + \sum_{c=1}^{C} \ln \sigma_{nc}^2 \quad (5)$$

where subscript c represents a component of a C-dimensional feature vector, and σ_{nc}^2, the cth diagonal component of the covariance matrix

$$\Sigma_n.$$

Further, if $\sigma_{nc}^2 = 1$ (the covariance matrix is a unit matrix) here, equation (3) can be simplified into the following equation.

$$d_t(n) = ||a_t - \hat{a}_t(n)||^2 \quad (6)$$

This equation (6) is the scale of distance used in NPM's according to the prior art, in which differences in the extent of fluctuation of the prediction residual from component to component of the feature vector are not taken into account.

The distance (global distance) D between input speech and an NPM is given by the following equation as the cumulation of local distances:

$$D = \min_{\{n_t\}} \sum_{t=1}^{T} d_t(n_t) \quad (7)$$

where n_t is the identification number of the NPM which performs prediction of the feature vector of input speech at time t. The minimization in equation (7) means the selection, out of possible trajectories $n_1, n_2, \ldots, n_t, \ldots n_T$ (possible status transitions on the finite status transition network) between input speech and an NPM, of what will minimize the global distance (accumulated prediction residual) D. Where the skipless left-to-right pattern shown in FIG. 2 is to be used as the NPM, n_t should satisfy the following constraints:

$$n_1 = 1 \quad (8)$$

$$n_T = N \quad (9)$$

$$n_t = n_{t-1} \text{ or } n_{t-1} + 1 \ (1 < t \leqq T) \quad (10)$$

where T is the length of the feature vector time series patterns of input speech signals, and N, the number of NPM statuses (the identification number of the final status). Under these constraints, the problem of minimization can be solved by dynamic programming (DP) using the following recursion formula (for details on DP, reference may be made to H. Sakoe and S. Chiba, "Dynamic Programming Algorithm Optimization for Spoken Word Recognition", IEEE Transaction on Acoustics, Speech, and Signal Processing, ASSP-26 (1), February 1978, pp. 43-49):

$$g_t(n) = d_t(n) + \min \begin{Bmatrix} g_{t-1}(n) \\ g_{t-1}(n-1) \end{Bmatrix} \quad (11)$$

where $g_t(n)$ is the partial sum of local distances $d_t(n)$, and the global distance D is given by the following equation.

$$D = g_T(N)$$

By tracing back the results, the optimal trajectory $\{n_t^*\}$ to minimize the accumulated prediction residual can be obtained. This information is used in the training algorithm to be described below. In recognizing continuous speech or the like, the word sequence of the recognition result can be identified from this information. FIG. 3 illustrates an outline of the recognition algorithm so far described.

Next will be described the training algorithm for automatically determining the parameters of NPM's (the weighting parameter for the neural network constituting the predictor and the parameter of the covariance matrix) by using known speech data. The purpose of training is to find out such model parameters as would minimize the aforementioned accumulated prediction residual for the speech data for training use. This can be formulated as follows as a problem of minimization having as its evaluative function the total D_{total} of the accumulated prediction residual for the whole training speech data:

$$D_{total} = \sum_{m=1}^{M} D(m) \quad (13)$$

where M is the total number of the sets of training data, and D(m), the accumulated prediction residual for the m-th set of training data. D(m) can be calculated by the algorithm using DP, formulated in the preceding section. The evaluative function D_{total} can be minimized in the optimal manner by the iterative algorithm stated below, combining DP and back-propagation (BP) (for more details on BP, reference may be made to R.P. Lippmann, "An Introduction to Computing with Neural Nets", IEEE ASSP Magazine, 3, 1987, pp. 4-22).

Step 1: Initialize all the NPM parameters (including the inter-unit coupling coefficient matrixes, threshold vectors and covariance matrixes of all the MLP predictors)

Step 2: m = 1

Step 3: Calculate the accumulated prediction residual D(m) for the mth set of training data by DP. Seek for the optimal trajectory $\{n_t^*\}$ by back-tracking.

Step 4: t = 1

Step 5: Assign a desirable output a_t to the output output $\hat{a}_t(n_t^*)$ of the n_t^*th MLP predictor of reference patterns, and calculate the correction quantity of each parameter by BP.

Step 6: t = t + 1

Step 7: If t is not greater than T_m (T_m is the the number of frames of the mth set of training data, return to step 5.

Step 8: m = m + 1

Step 9: If m is not greater than M, return to step 3.

Step 10: Update all the NPM parameters according to the correction quantities calculated at step 5.

Step 11: If the conditions of convergence are not satisfied, return to step 2.

While parameter corrections by BP in the foregoing algorithm use the determined steepest descent method by which all the corrections are done collectively at step 10, the corrections can as well be consecutively accomplished by the random steepest descent method. Regarding the conditions of convergence at step 11, the convergence is deemed to have been achieved when, for instance, the decrement of the evaluative function D_{total} drops below a certain level.

When a covariance matrix is introduced into the scale of local distances, the quantity of back-propagation error by BP requires the following correction (incidentally, the nondiagonal terms of the covariance matrix are small enough to ignore). The amount of the back-propagation error of the c-th unit of the output layer of an MLP predictor in the n_t^*-th status ($\sigma_{tv}(n_t(n_t^*))$ is:

$$\delta_{tv}(n_i^*) = \frac{a_{tv} - \hat{a}_{tv}(n_t^*)}{\sigma^2_{n_t^* v}} \quad (14)$$

This differs from the amount of back-propagation error taking account of no covariance matrix by the reciprocal of variance. The estimated value of the covariance matrix is so determined as to minimize the evaluative function D_{total}. Thus from the following optimizing condition:

$$\frac{\partial D_{total}}{\partial \sigma^2_{nv}} = 0 \quad (15)$$

is derived the following estimation formula (revaluation formula for use at step 10) of the covariance matrix:

$$\sigma^2_{nv} = \frac{\sum_{m=1}^{M} \sum_{t=1}^{T_m} \delta_{nn_t^*} (a_{tv} - \hat{a}_{tv}(n_i^*))^2}{\sum_{m=1}^{M} \sum_{t=1}^{T_m} \delta_{nn_t^*}} \quad (16)$$

where T_m is the number of frames of the m-th set of training data, $\sigma_{nn_t^*}$, a Kronecker delta.

$$\delta_{mn} = \begin{cases} 1 \text{ if } m = n \\ 0 \text{ otherwise} \end{cases} \quad (17)$$

The convergence of the foregoing iterative training algorithm can be proven in the following way. Thus, the value $D^{(k)}_{total}$ of the evaluative function in the k-th iteration before the parameter correction (immediately before step 10) is the sum of prediction residuals accumulated according to the optimal (accumulated prediction residual minimizing) trajectory $\{n_i^*\}$ determined by DP for each set of training data. The sum of prediction residuals accumulated according to the same trajectory after the parameter correction at step 10 is represented by $D^{(k,BP)}_{total}$. Whereas the parameter correction by BP here is so accomplished as to reduce the squared error in the output layer of each MLP predictor, in the case of NPM, where this squared error is identical with the prediction residual, the accumulated prediction residual is reduced without fail by the parameter correction. (Revaluation of the covariance matrix is considered together with BP.)

$$D^{(k,BP)}_{total} \leqq D^{(k)}_{total} \quad (18)$$

However, when the parameters are corrected by BP, the optimality of the optimal trajectory achieved at step 3 is lost. Therefore, the optimal trajectory is sought for the model parameters corrected by DP in the k+1-th iteration. Since DP gives the optimal trajectory to minimize the accumulated predictor residual:

$$D^{(k+1)}_{total} \leqq D^{(k,BP)}_{total} \quad (19)$$

formulas 18 and 19 eventually indicate that the evaluative function is monotonously reduced by iteration.

$$D^{(k+1)}_{total} \leqq D^{(k)}_{total} \quad (20)$$

Qualitatively it can be understood that the reason why this iterative algorithm converges is that DP and BP are minimization methods for the same evaluative function (the accumulated sum of prediction residuals) and they are consecutively applied.

DETAILED DESCRIPTION

The present invention will be described in further detail below with reference to FIGS. 4 to 6 which are the flow charts of recognition by the speech recognition method according to the invention, FIG. 7 which is a flow chart of reference pattern learning by the speech recognition method according to the invention and FIG. 8 which is a block diagram of a speech recognition apparatus according to the invention.

A speech input unit 8101 in FIG. 8, consisting of a microphone, an amplifier and an analog-to-digital (A/D) converter among other things, digitalizes speech signals representing speech sounds uttered by the user and supplies them to the following acoustic analyzer 8102. The acoustic analyzer 8102 subjects these digitalized speech signals to spectral analysis by FFT or the like, and converts them into a time series pattern of feature vectors. The spectral analysis can be accomplished by linear predictive coding (LPC) or the cepstrum method

besides FFT.

A reference pattern storage section 8103 stores the parameters of the reference pattern models of all the word categories which are the objects of recognition. If, for instance, 10 numerals are to be recognized, the parameters of the reference pattern model of each of the numerals from 0 through 9 are stored. The reference pattern model of each category here is a finite status transition network each of whose statuses is accompanied by an MLP predictor.

What are stored in the storage section 8103 are the parameters of the MLP predictors of different statuses and the covariance matrixes of the respective statuses. Where MLP predictors each having one hidden layer, as shown in FIG. 1, are used, the parameters are τ_F inter-unit coupling coefficient matrixes W_1^F, ..., $W_{\tau_F}^F$ (each matrix consists of H rows by C columns, where C is the number of hidden layer units and C is the number of dimensions of the feature vector) for forward prediction, τ_B inter-unit coupling coefficient matrixes W_1^B, ..., $W_{\tau_B}^B$ (each matrix consists of H rows by C columns) for backward prediction, an inter-unit coupling coefficient matrix W (consisting of C rows by H columns), the threshold vector Θ_1, of the hidden layer (H-dimensional vector), and the threshold vector Θ_0 of the output layer (C-dimensional vector). Each covariance matrix is a symmetric one of C rows by C columns, and the number of independent components is C(C + 1)/2.

A distance calculator 8104 calculates the distance between the feature vector time series pattern given from the acoustic analyzer 8102 and the reference pattern model of each of the aforementioned categories, and supplies the result of calculation to a recognition result output section 8105. The distance between the feature vector time series pattern of input speech and the reference pattern model is the global distance D defined by the above cited equation (7). The method of calculation is given by the flow from steps 401 to 414 in FIG. 4. In FIG. 4, t is a variable representing the time axis of the feature vector time series pattern of input speech, and takes one of integral values from 1 through T; s is a variable representing the category of the objects of recognition, and takes one of integral values from 1 through S - where the objects of recognition are 10 numerals, s = 10; n is a variable representing the status of the reference pattern model of each category, and takes one of integral values from 1 through $N^{(S)}$ ($N^{(S)}$ is the number of status of the reference pattern model of category s); $d_t^{(S)}(n)$ is a variable for storing the local distance between the n-th status of category s and the feature vector a_t of input speech at time t; and $g_t^{(S)}(n)$ is a variable for storing the accumulated prediction residual of the n-th status of category s at time t.

At step 401, the variables are initialized in order to calculate the global distance D of equation (7) by DP. Referring to FIG. 5 illustrating this step 401 in detail, a counter is initialized at steps 501 to 503. At step 504, the storage areas for the local distance $d_{st}(n)$ and the accumulated prediction residual $g_{st}(n)$ are initialized. At steps 505 to 510, the increment and conditions of the counter are judged, and the initialization at step 504 is applied with respect to all the values of s, t and n. Then, at steps 511 to 514, the value at the start point of the accumulated prediction residual of each category s is set.

Next, referring to FIG. 6 illustrating step 405 for calculating the local distance in detail, at step 601, an MLP predictor accompanying the n-th state of category s calculates the predicted vector $\hat{a}_t$ to be compared with the input speech feature vector a_t at time t. In FIG. 9 which shows in further detail this calculation represented by the above cited equations (1) and (2), X is a scalar variable; Y, the arrangement in the H-dimension (Y(h) is the h-th element); Z, the arrangement in the C-dimension (Z(c) is the c-th element); H, the number of hidden layer units; and C, the number of dimensions of the feature vector. Further, $(\Theta_1)_h$ at step 9202 is the h-th component of the threshold vector Θ_1^1; $(W_S^F)_{hc}$ at step 9205 is the element on the h-th row and the c-th column of the coupling coefficient matrix W_S^F; and τ_B represent the numbers of feature vectors of input speech used for the aforementioned forward prediction and backward prediction, respectively, $\tau_F = 2$ and $\tau_B = 1$ being used, to be specific. The calculation shown in FIG. 9 gives the predicted vector $\hat{a}_t$ as a vector arrangement Z having C components.

Next, referring further to FIG. 6, at step 602 is calculated the distance $d_t^{(S)}(n)$ from the input speech feature vector a_t at time t and the predicted vector $\hat{a}_t$ by the MLP predictor accompanying the n-th status of category , calculated at step 601. At step 10301 of FIG. 10 which shows in further detail this calculation represented by the above cited equation (3),

$$|\Sigma_n^{(s)}|$$

is the determinant of the covariance matrix in the n-th state of category s. The method to calculate the determinant is not explained here as it is evident from elementary linear algebra. At step 10301, the natural logarithm

of the determinant of the covariance matrix is substituted for a variable X. Variables Y and X in FIG. 10 are both arranged in the C-dimension. At step 10305, $(a_t)_{c2}$ is the c -th component of the input speech feature vector a_t, and $(a_t^{(S)}(n))_{c2}$ is the predicted vector by the MLP predictor accompanying the n-th state of category s, both calculated at step 601. At step 10306,

$$(\Sigma_n^{(s)-1})_{c1c2}$$

is the component on the c_1-th row and the c_2-th column of the inverse matrix of the covariance matrix

$$\Sigma_n^{(s)}.$$

The method to calculate the inverse matrix is not explained here as it is evident from elementary linear algebra. Processing illustrated in FIG. 10 stores the value of the local distance $d_t^{(S)}(n)$ into the variable X.

By the processing up to step 414 in FIG. 4, the global distance D between the featu vector time series pattern of input speech and the reference pattern model is calculated. At this time, the global distance D between the above mentioned patterns of category s is obtained as the accumulated prediction residual $g_T(N^{(S)})$ of the final status $N^{(S)}$ of each reference pattern model at time T (the terminal point of one time series pattern).

The recognition result output section 8105 selects the shortest of the distances between the feature vector time series pattern of input speech and the reference pattern model of each category given from the distance calculator 8104, and supplies its category name as the recognition result. Its specific processing is step 415 in FIG. 4.

A training speech database storage section 8106 stores the speech data of all the word/syllable categories which are the objects of recognition, i.e. the time series patterns of feature vectors corresponding to each category.

A reference pattern corrector 8107 calculates the quantities of correction required for the parameters of the reference pattern model of each category read in from the reference pattern storage section 8103 on the basis of training speech data from the training speech database storage section 8106, and corrects the aforementioned parameters corresponding to the reference patterns stored in the reference pattern storage section 8103.

Referring to FIG. 7 illustrating that signal processing, the parameters of the reference pattern models of all the categories (including the inter-unit coupling coefficients of MLP predictors in different statuses, threshold vectors and covariance matrixes) are initialized with random numbers at step 701. This processing is illustrated in detail in FIG. 11. Here, s is a variable representing the category of the object of recognition, and takes, where 10 numerals are to be recognized, one of integral values from 1 through 10; and n is a variable representing the n-th status of the reference pattern model of the s-th category, and takes one of integral values from 1 through $N^{(S)}$. At step 11406, $(W_t^F)_{hc}$ is the element on the h-th row and the c-th column of the t-th inter-unit coupling matrix W_t^F for forward prediction of the MLP predictor accompanying the n-th status of the reference pattern model of the s-th category. Here, "random" denotes random numbers, which specifically are uniform random numbers ranging from -0.3 to 0.3. Similarly at the following steps 11416, 11424, 11429 and 11432, the object parameter is supposed to be the parameter of the n-th status of the reference pattern model of the s-th category. At step 11439 are initialized variables D1 and D2, to be subsequently used for convergence decision at step 712, and another variable P.

At step 704, the optimal trajectories between the m-th set of training data of the s-th category (the feature vector time series of $T_m^{(S)}$ in length) and the reference pattern models of the s-th category are figured out. Details of this processing are shown in FIG. 12. The initializing section 2 of step 12501 results from the fixing of the variable s representing the category to be processed in the initialization given in FIG. 4 (step 401). More specifically, this processing is achieved by eliminating steps 502, 507, 508, 511, 513 and 514 of FIG. 5. The distance calculating section 2 of step 12502 results from the fixing of the variable s representing the category to be processed in the distance calculating process from steps 402 through 414 given in FIG. 4. More specifically, this processing is accomplished by removing steps 403, 411 and 412. For subsequent convergence decision, the accumulated prediction residual $g_T^{(S)}(N^{(S)})$ calculated by the processing up to step 414 is added in advance

to the variable D1 for convergence decision ($D1 = D1 + g_T^{(S)}(N^{(S)})$. The optimal trajectories $n_1, ..., n_T$ are obtained by the processing from steps 12503 through 12510.

Referring again to FIG. 7, at step 706, the parameters of the MLP predictor accompanying the n_i^*-th status, given correspondence at step 704 to the feature vector a_t of the m-th set of training data of the s-th category at time t, are corrected by back propagation. Details of this processing are shown in FIG. 13. At step 13602 is calculated the predicted vector for the feature vector a_t at time t. This process is shown in FIG. 9 (referred to above). In the processing at the following steps 13603 through 13634, Y is an H-dimensional arrangement representing the output of the hidden layer unit calculated at step 13602; ΔZ, a C-dimensional arrangement representing the error regarding the output layer unit; ΔY, an H-dimensional arrangement representing the error regarding the hidden layer unit; and ε, a learning coefficient given in advance (specifically taking the value of 0.1 or the like). Here, the nondiagonal terms of the covariance matrix are negligibly small, and the matrix accordingly is treated as a diagonal one, whose c-th diagonal component is

$$(\Sigma_n)_{cc}$$

at step 13607. By the processing shown in FIG. 13, the parameters of the n_i^*-th MLP predictor of the s-th category are so corrected as to reduce the prediction residuals. By the processing from steps 703 through 710, the above described corrective training is applied to all the sets of training data of the s-th category.

At step 711, a new covariance matrix is calculated on the basis of equation (16) cited above. The process at step 711 is illustrated in FIG. 14, wherein variables X and Y are arrangements of $N^{(S)}$ rows and C columns, respectively. The optimal trajectory calculation at step 14709 is processed in the same manner as step 704, as shown in detail in FIG. 12. Processing by the predicted vector calculating section at step 14712 is the same as step 13602, whose details are shown in FIG. 9. The sign ($\hat{a}_t(n)$) at step 17414 denotes the c-th component of the predicted vector calculated at step 14712, and

$$(\Sigma_n)_{cc}$$

at step 14724, the c-th diagonal component of the the covariance matrix of the n-th status.

Conversion decision at step 712 recognizes convergence if the calculated varying rate of the variable D1 (the accumulated prediction residual for all the sets of training data) is found smaller than a threshold given in advance. The processing is shown in FIG. 15. At step 15802, the absolute value of the varying rate of the accumulated prediction residual for all the sets of training data is compared with a threshold T_h given in advance (actually 0.001 or the like). By the processing at these steps 701 through 417, iterative training is carried out for all the sets of training data to give the optimal model parameters.

As hitherto described, the speech recognition method according to the present invention is characteristic in that finite status transition networks of the left-to-right pattern accompanied by MLP predictors using neural networks compose NPM's, which are reference pattern models involving both forward and backward prediction, the local distances between these NPM's and the time series patterns of input speech feature vectors are calculated by DP matching, and covariance matrixes are introduced into this calculation of local distances. As a result, the speech recognition method according to the invention is adaptable to the speech of any unspecified speaker and makes possible speech recognition, in particular continuous speech recognition, at a high rate of recognition.

Although the invention has been described with reference to a specific embodiment (or specific embodiments), this description is not meant to be construed in a limiting sense. Various modifications of the disclosed embodiment, as well as other embodiments of the invention, will become apparent to persons skilled in the art upon reference to the description of the invention. It is therefore contemplated that the appended claims will cover any such modifications or embodiments as fall within the true scope of the invention.

Claims

1. A pattern recognition method for recognizing syllables and sound elements on the basis of comparison of input time series patterns expressed as the feature vectors of the syllables and sound elements with ref-

erence pattern models using a finite status transition network, wherein each status of said finite status transition network has a predictor for calculating a predicted vector at time t from a plurality of feature vectors of said input time series patterns at time t-1 and before and a plurality of feature vectors at time t+1 and after; and a prediction residual determined by said input feature vector, said predicted feature vector by the predictor of said n-th status at time t, corresponding to said feature vector, and a covariance matrix accompanying said nth status is used as the local distance between the feature vector of said input time series patterns at time t, i.e. the input feature vector, and the n-th status of said finite status transition network.

2. A speech recognition method, as claimed in Claim 1, wherein correspondences between said input time series patterns and said predictors are so figured out by dynamic programming as to minimize the accumulated value of said local distances along said status transitions, and said accumulated value is used as the distance between said input time series patterns and said reference pattern models.

3. A speech recognition method, as claimed in Claim 2, wherein said accumulated value is calculated for every category of words, and the category having the smallest of said accumulated values which have been calculated is used as the recognition output.

4. A speech recognition method, as claimed in Claim 2 or 3, wherein initial values are set for the parameters of said predictor and said covariance matrix accompanying each status of said finite status transition network, said distance between said input time series pattern for the learning purpose, said category of which is known, and said reference pattern model corresponding to the same category as said known category is calculated; the parameters of said predictor and said covariance matrix of each state are iteratively corrected in the direction of reducing said distance; and said reference pattern model said distance for which satisfies predetermined conditions of convergence is thereby obtained.

Fig. 1

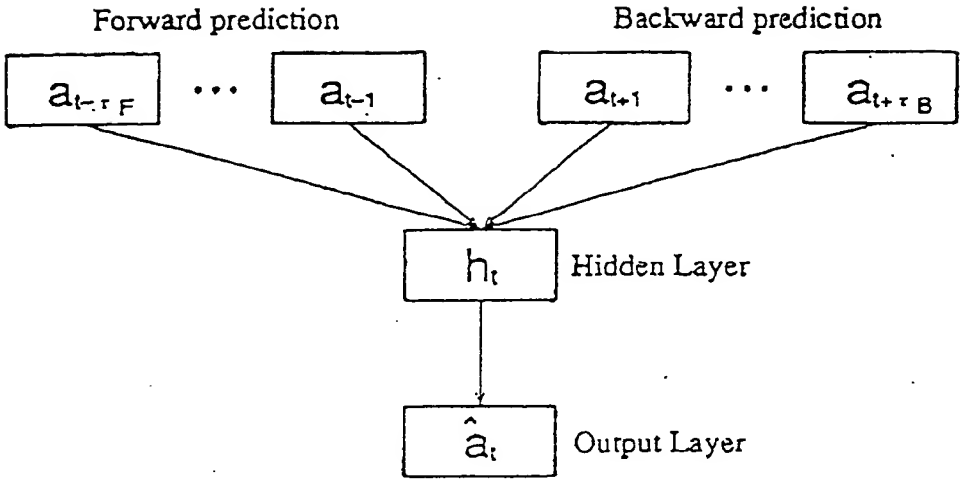

Forward prediction
Backward prediction
$a_{t-\tau F}$
a_{t-1}
a_{t+1}
$a_{t+\tau B}$
h_t
Hidden Layer
$\hat{a}_t$
Output Layer

Fig. 2

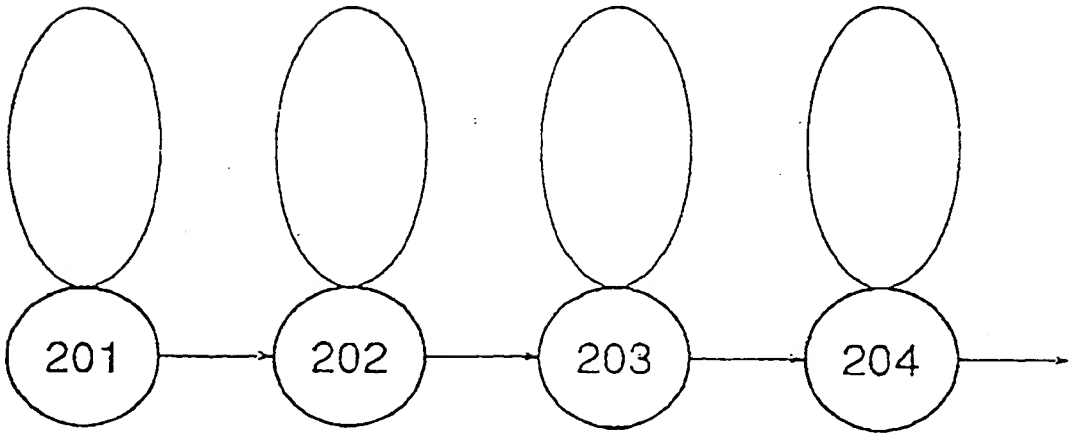

201
202
203
204

Fig. 3

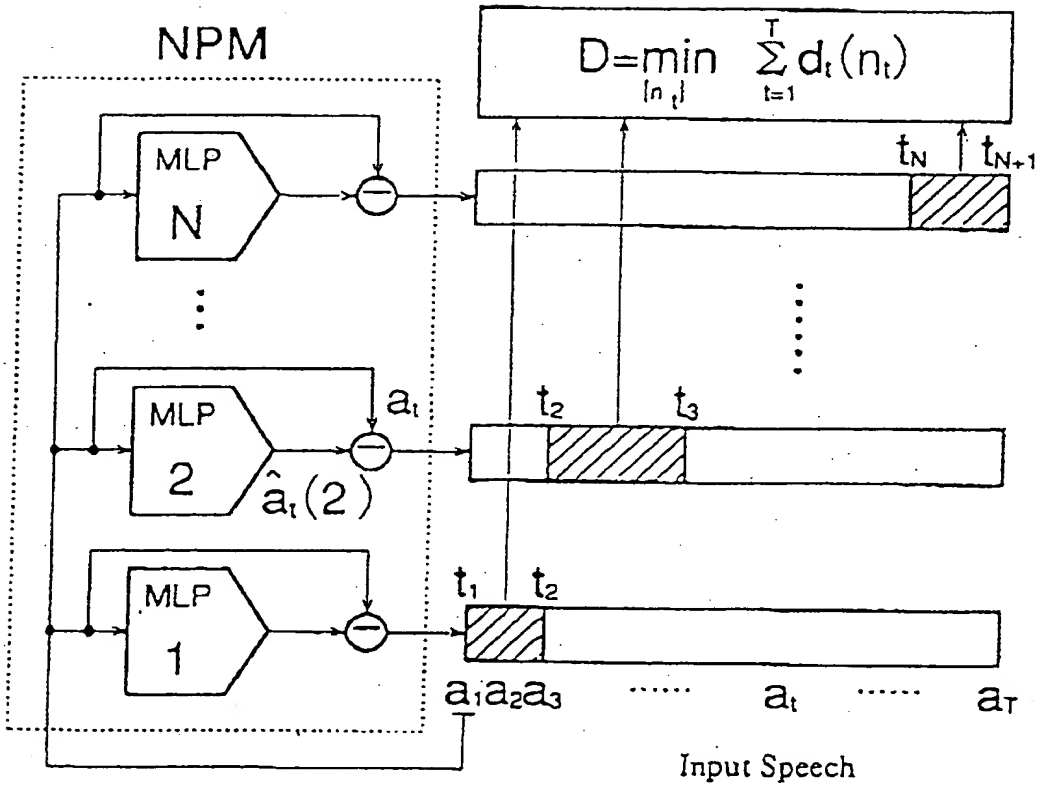

NPM
$D=\min_{[n_t]} \sum_{t=1}^{T} d_t(n_t)$
MLP
N
MLP
2
MLP
1
a_t
$\hat{a}_t(2)$
t_1
t_2
t_3
t_N
t_{N+1}
$a_1 a_2 a_3$ ..... a_t ..... a_T
Input Speech

Fig. 4

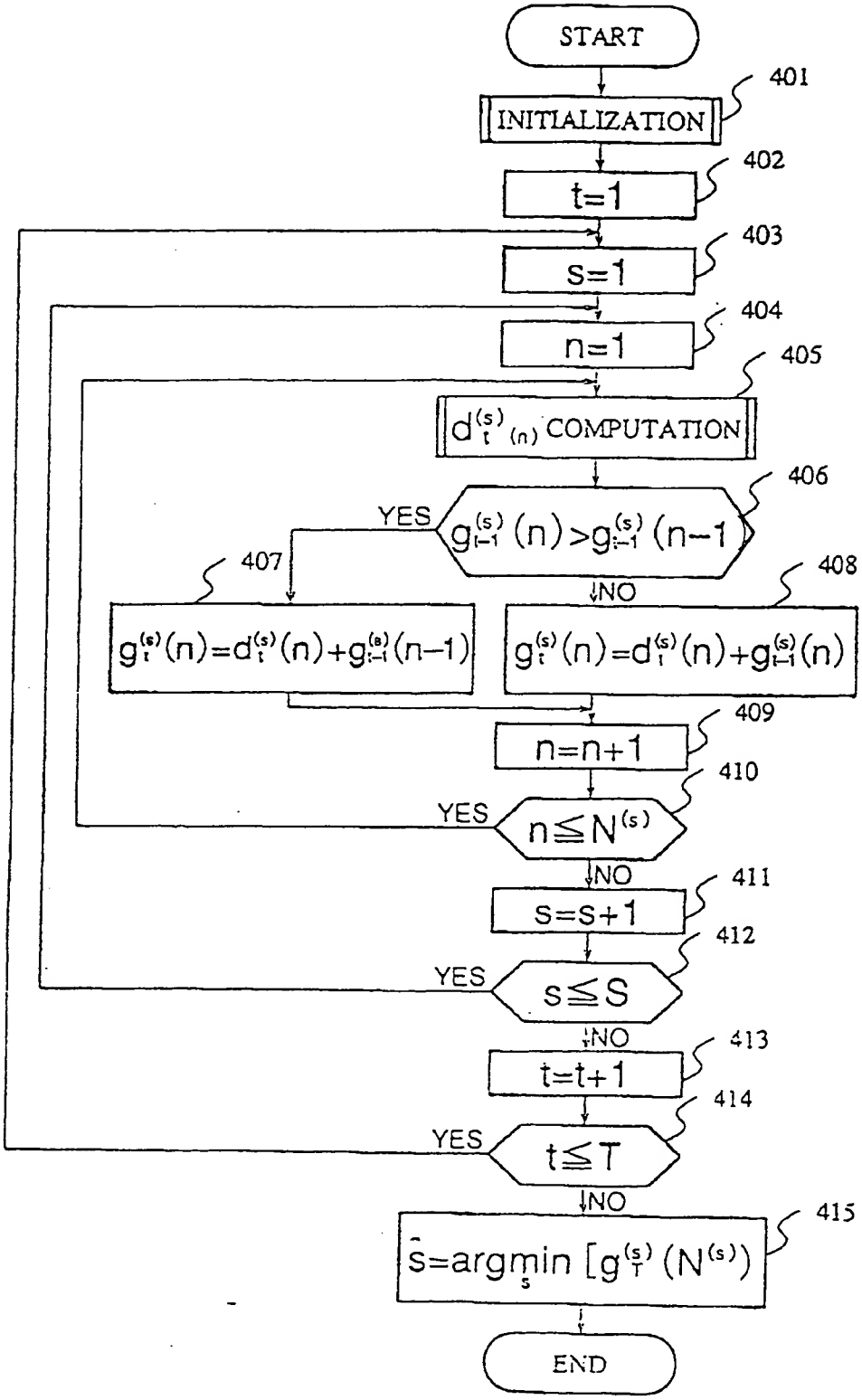

START
401
INITIALIZATION
402
t=1
403
s=1
404
n=1
405
$d_t^{(s)}(n)$ COMPUTATION
406
YES
$g_{t-1}^{(s)}(n) > g_{t-1}^{(s)}(n-1)$
407
NO
408
$g_t^{(s)}(n) = d_t^{(s)}(n) + g_{t-1}^{(s)}(n-1)$
$g_t^{(s)}(n) = d_t^{(s)}(n) + g_{t-1}^{(s)}(n)$
409
n=n+1
410
YES
$n \leqq N^{(s)}$
NO
411
s=s+1
412
YES
$s \leqq S$
NO
413
t=t+1
414
YES
$t \leqq T$
NO
415
$\hat{s} = \underset{s}{\text{argmin}} [g_T^{(s)}(N^{(s)})]$
END

Fig. 5

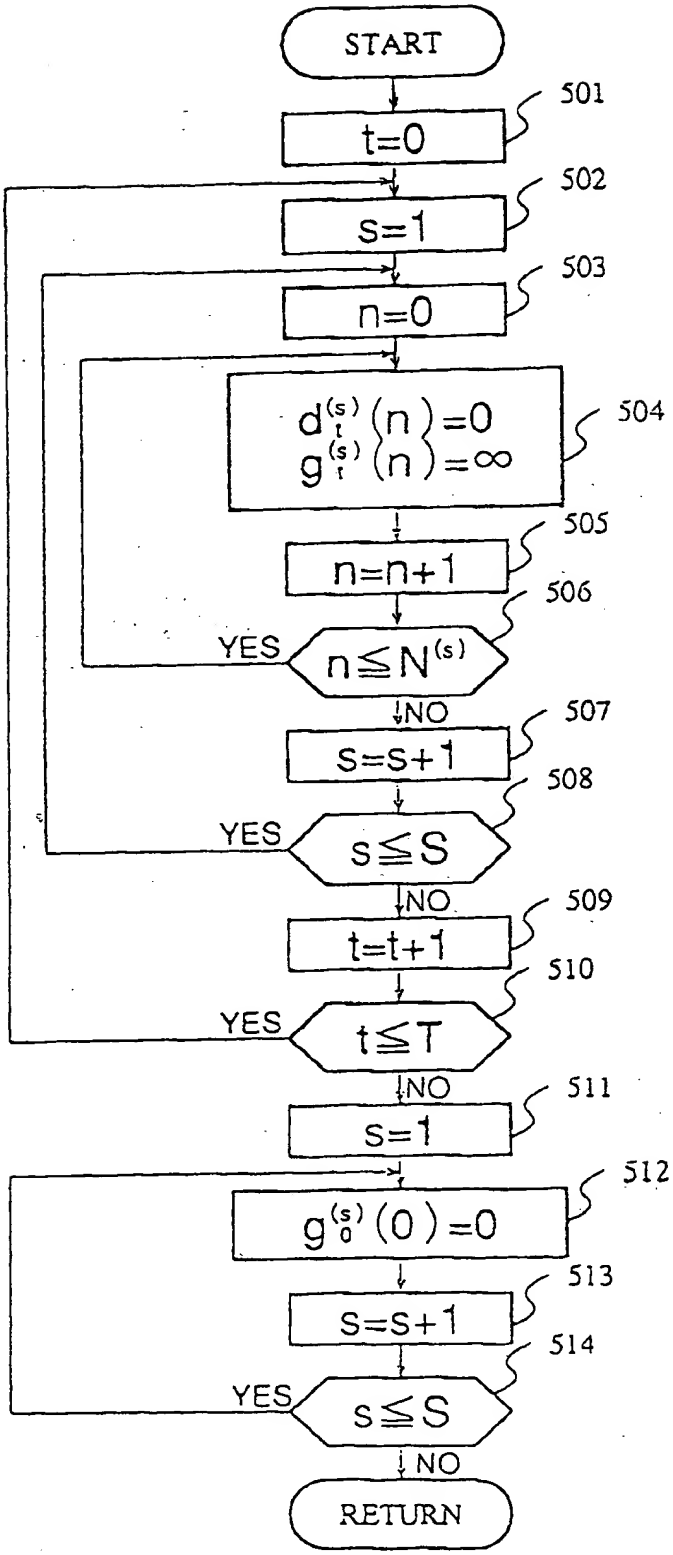

START
501
t=0
502
s=1
503
n=0
504
$d_t^{(s)}(n)=0$
$g_t^{(s)}(n)=\infty$
505
n=n+1
506
YES
$n \leqq N^{(s)}$
NO
507
s=s+1
508
YES
$s \leqq S$
NO
509
t=t+1
510
YES
$t \leqq T$
NO
511
s=1
512
$g_0^{(s)}(0)=0$
513
s=s+1
514
YES
$s \leqq S$
NO
RETURN

Fig. 6

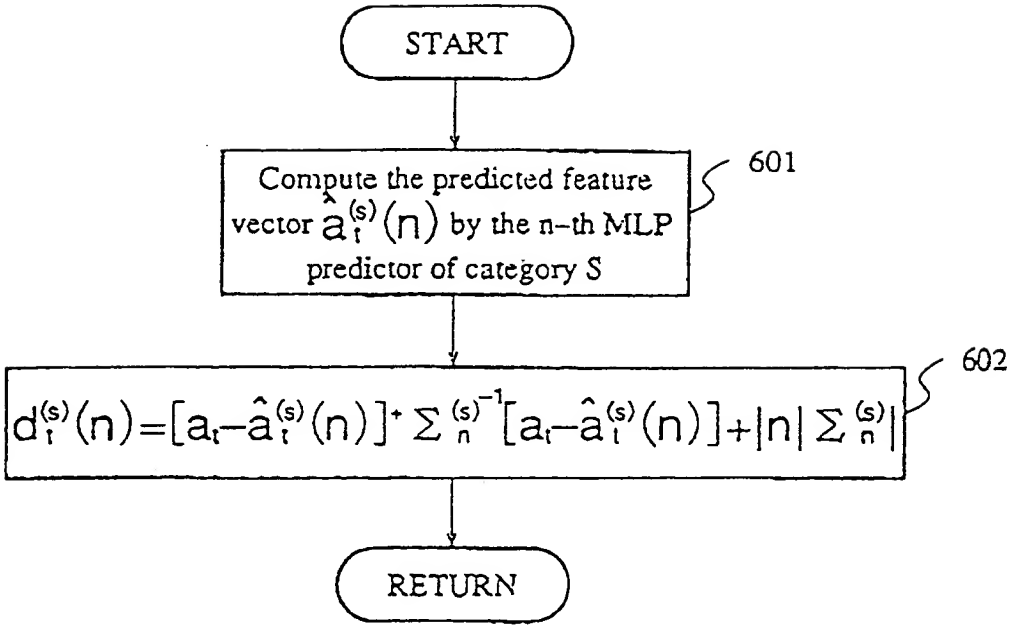

START
Compute the predicted feature vector $\hat{a}_t^{(s)}(n)$ by the n-th MLP predictor of category S
601
$d_t^{(s)}(n)=[a_t-\hat{a}_t^{(s)}(n)]^+ \Sigma_n^{(s)-1}[a_t-\hat{a}_t^{(s)}(n)]+|n|\,\Sigma_n^{(s)}|$
602
RETURN

Fig.7

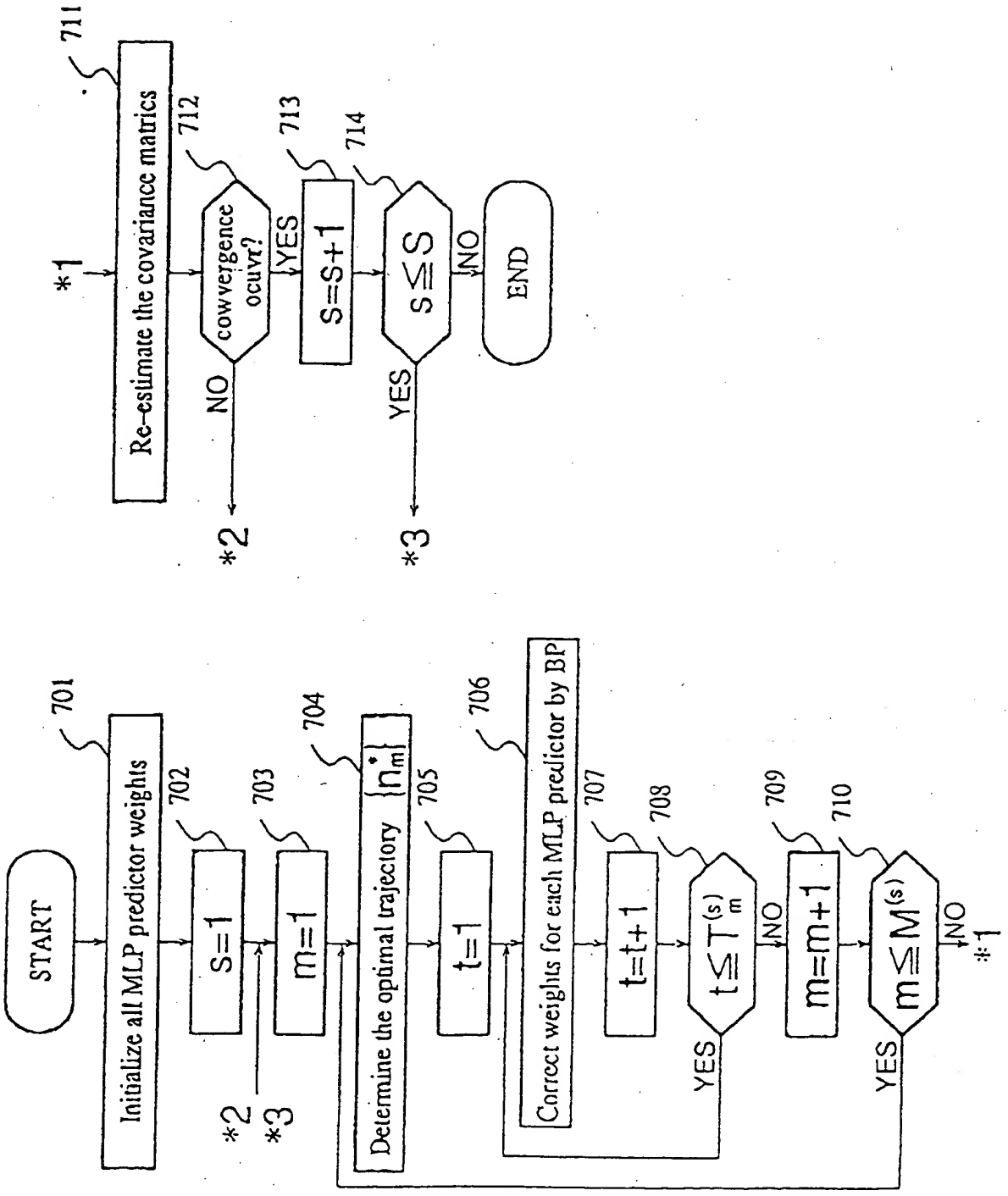

START
Initialize all MLP predictor weights
701
S=1
702
m=1
703
Determine the optimal trajectory $\{n_m^*\}$
704
t=1
705
Correct weights for each MLP predictor by BP
706
t=t+1
707
$t \leqq T_m^{(s)}$
708
YES
NO
m=m+1
709
$m \leqq M^{(s)}$
710
YES
NO
*1
*2
*3
Re-estimate the covariance matrics
711
convergence ocuur?
712
NO
YES
S=S+1
713
$S \leqq S$
714
YES
NO
END

Fig. 8

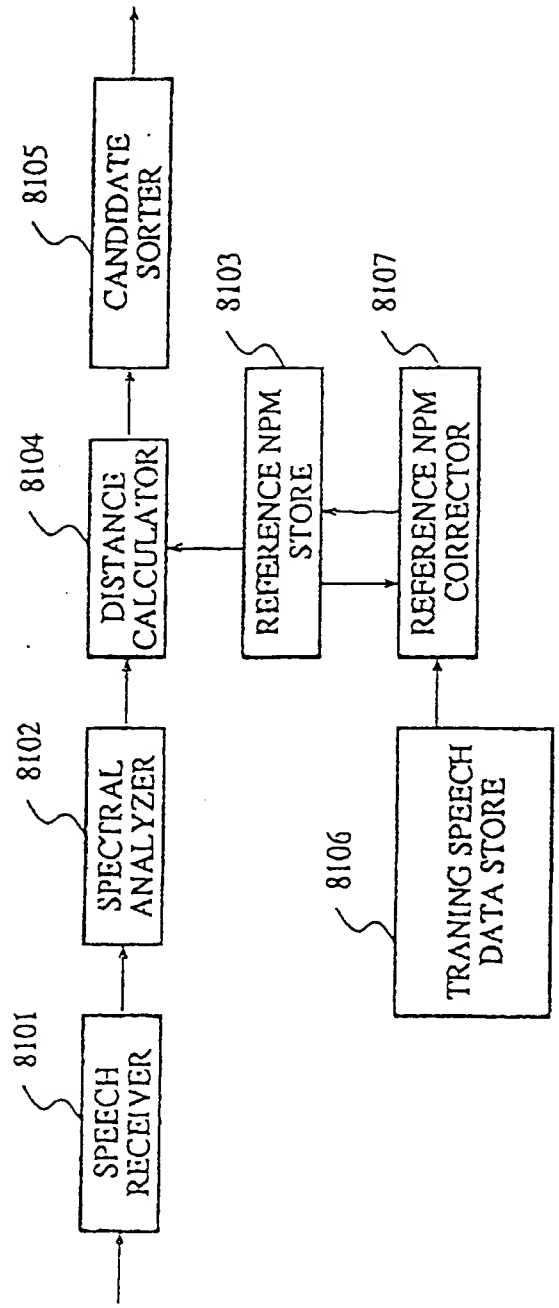

8101
SPEECH RECEIVER
8102
SPECTRAL ANALYZER
8104
DISTANCE CALCULATOR
8105
CANDIDATE SORTER
8103
REFERENCE NPM STORE
8107
REFERENCE NPM CORRECTOR
8106
TRANING SPEECH DATA STORE

Fig. 9

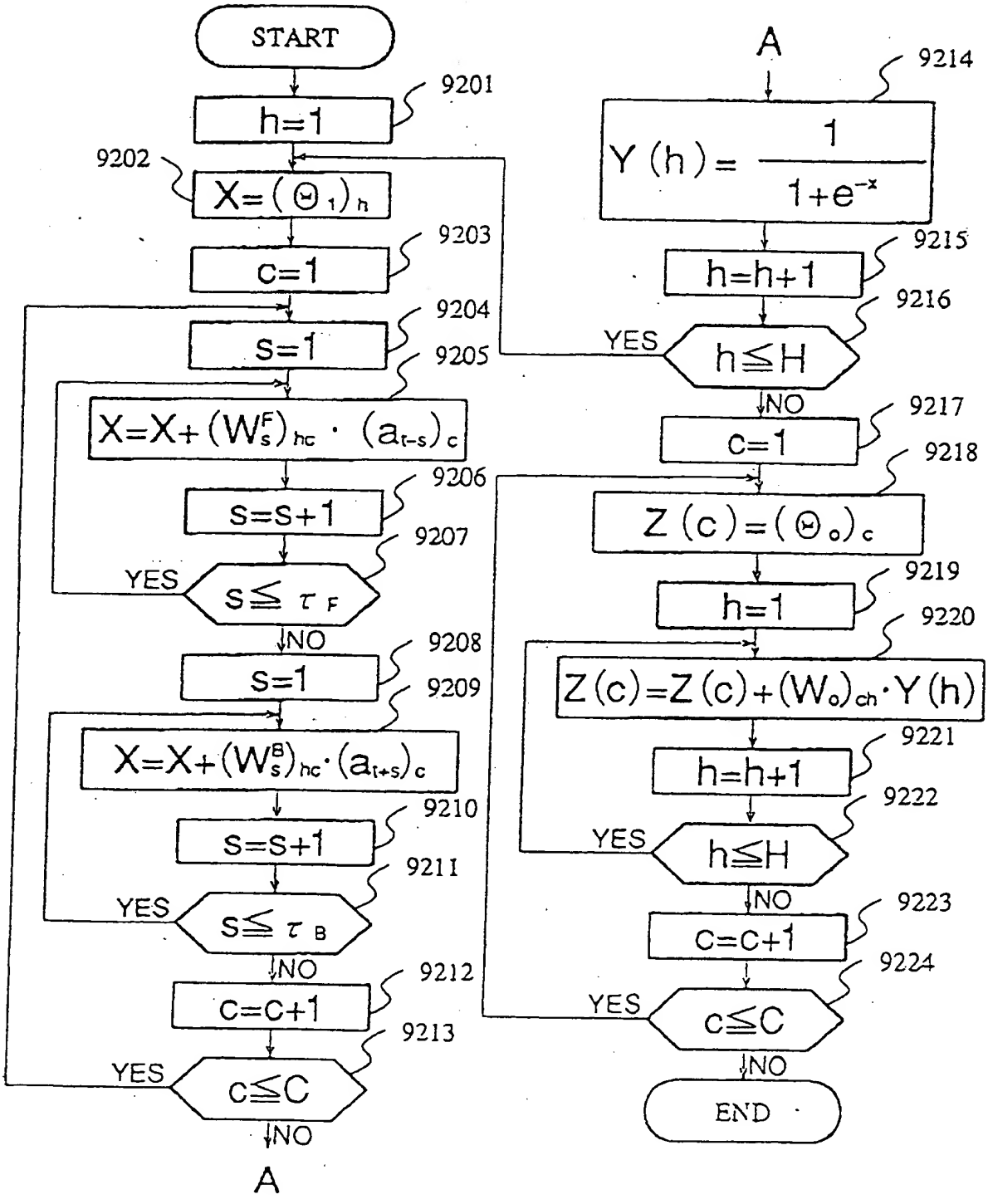

START
h=1
9201
9202
X=(Θ1)h
c=1
9203
s=1
9204
9205
X=X+(WsF)hc · (at-s)c
s=s+1
9206
9207
YES
s≦τF
NO
s=1
9208
9209
X=X+(WsB)hc·(at+s)c
s=s+1
9210
9211
YES
s≦τB
NO
c=c+1
9212
9213
YES
c≦C
NO
A
A
9214
Y(h)=1/(1+e^-x)
h=h+1
9215
9216
YES
h≦H
NO
c=1
9217
9218
Z(c)=(Θo)c
h=1
9219
9220
Z(c)=Z(c)+(Wo)ch·Y(h)
h=h+1
9221
9222
YES
h≦H
NO
c=c+1
9223
9224
YES
c≦C
NO
END

Fig:10

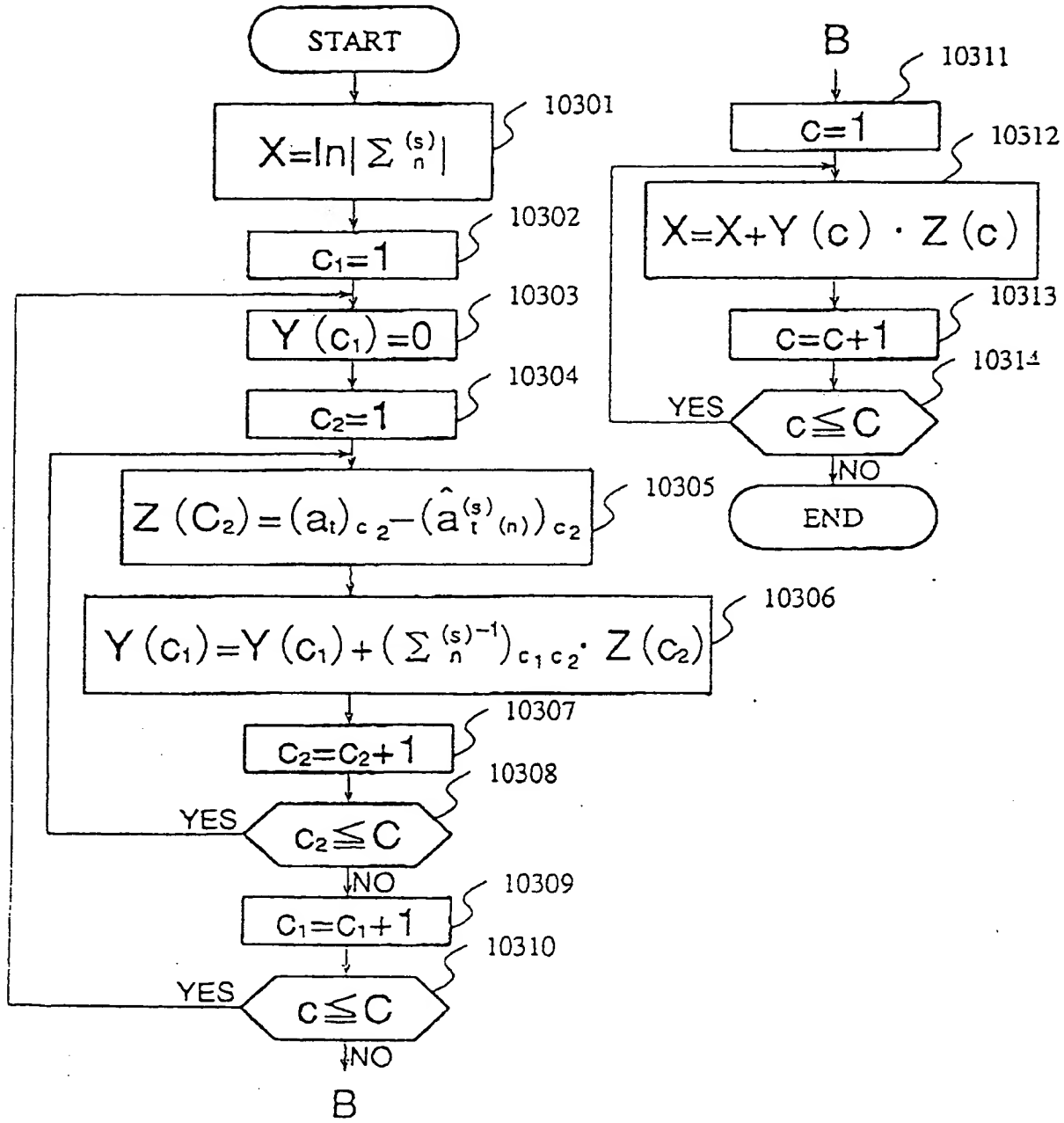

START
10301
$X=\ln|\Sigma_n^{(s)}|$
10302
$c_1=1$
10303
$Y(c_1)=0$
10304
$c_2=1$
10305
$Z(C_2)=(a_t)_{c_2}-(\hat{a}_{t(n)}^{(s)})_{c_2}$
10306
$Y(c_1)=Y(c_1)+(\Sigma_n^{(s)-1})_{c_1c_2}\cdot Z(c_2)$
10307
$c_2=c_2+1$
10308
YES
$c_2\leqq C$
NO
10309
$c_1=c_1+1$
10310
YES
$c\leqq C$
NO
B
B
10311
$c=1$
10312
$X=X+Y(c)\cdot Z(c)$
10313
$c=c+1$
10314
YES
$c\leqq C$
NO
END

Fig. 11

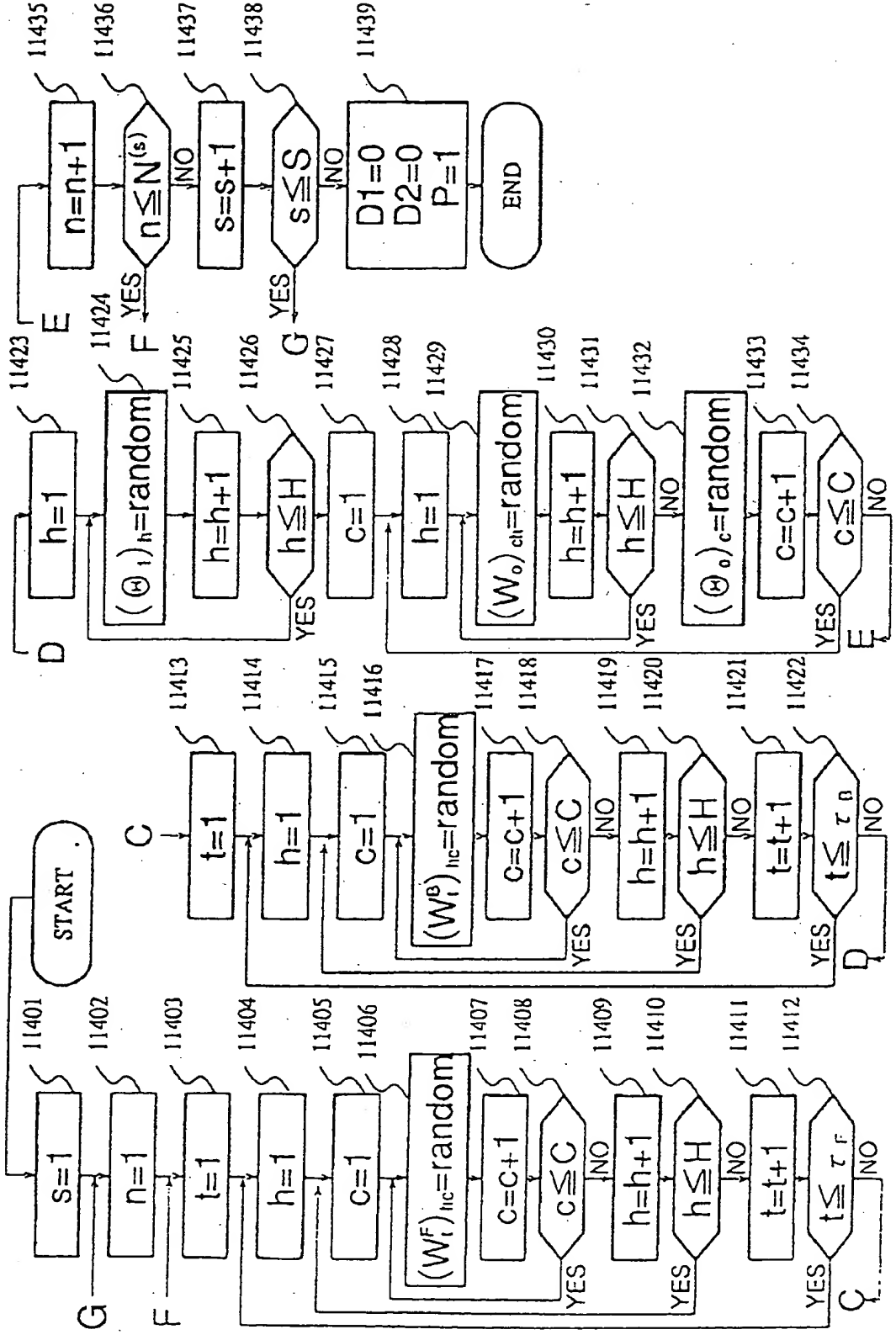

START
s=1
n=1
t=1
h=1
c=1
$(W_i^F)_{hc}$=random
c=c+1
c≦C
h=h+1
h≦H
t=t+1
t≦τ_F
t≦τ_B
$(W_i^B)_{hc}$=random
$(\Theta_i)_h$=random
$(W_o)_{ch}$=random
$(\Theta_o)_c$=random
n=n+1
n≦$N^{(s)}$
s=s+1
s≦S
D1=0
D2=0
P=1
END
YES
NO
A
B
C
D
E
F
G
11401
11402
11403
11404
11405
11406
11407
11408
11409
11410
11411
11412
11413
11414
11415
11416
11417
11418
11419
11420
11421
11422
11423
11424
11425
11426
11427
11428
11429
11430
11431
11432
11433
11434
11435
11436
11437
11438
11439

Fig. 12

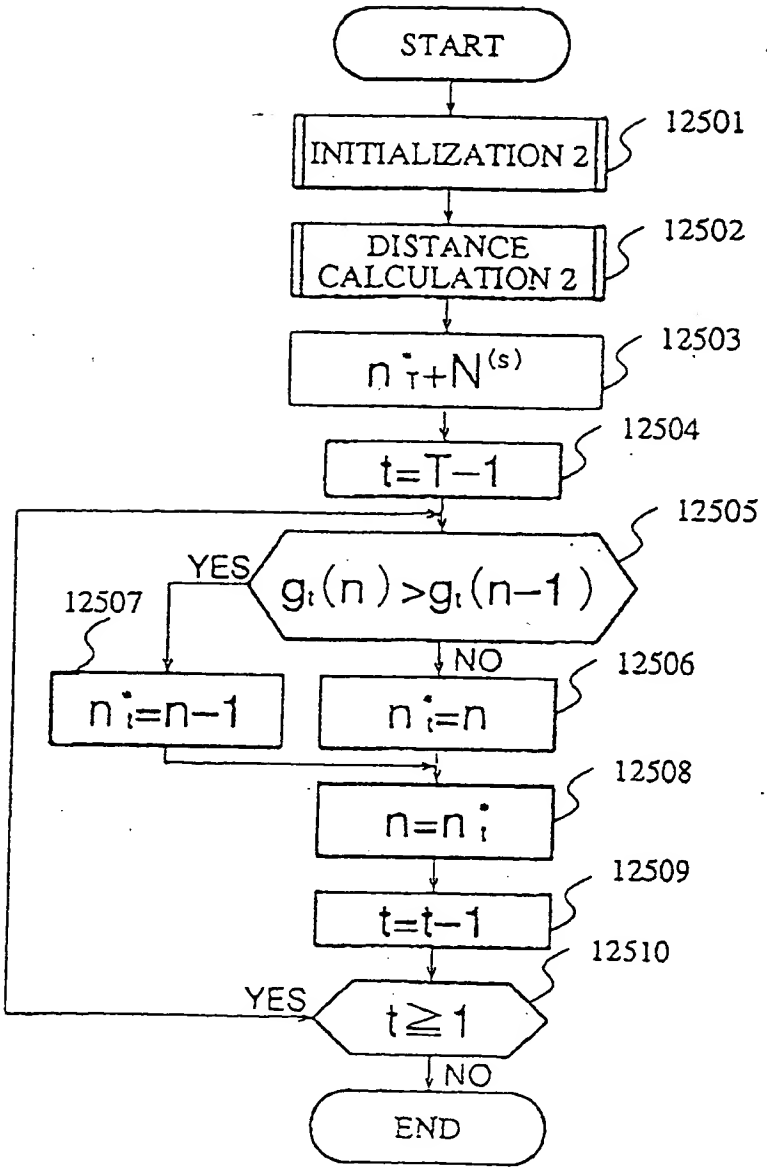

START
INITIALIZATION 2
12501
DISTANCE CALCULATION 2
12502
$n_T^* + N^{(s)}$
12503
t=T−1
12504
12505
YES
$g_t(n) > g_t(n-1)$
12507
NO
12506
$n_t^* = n-1$
$n_t^* = n$
12508
$n = n_t^*$
12509
t=t−1
12510
YES
t≧1
NO
END

Fig. 13

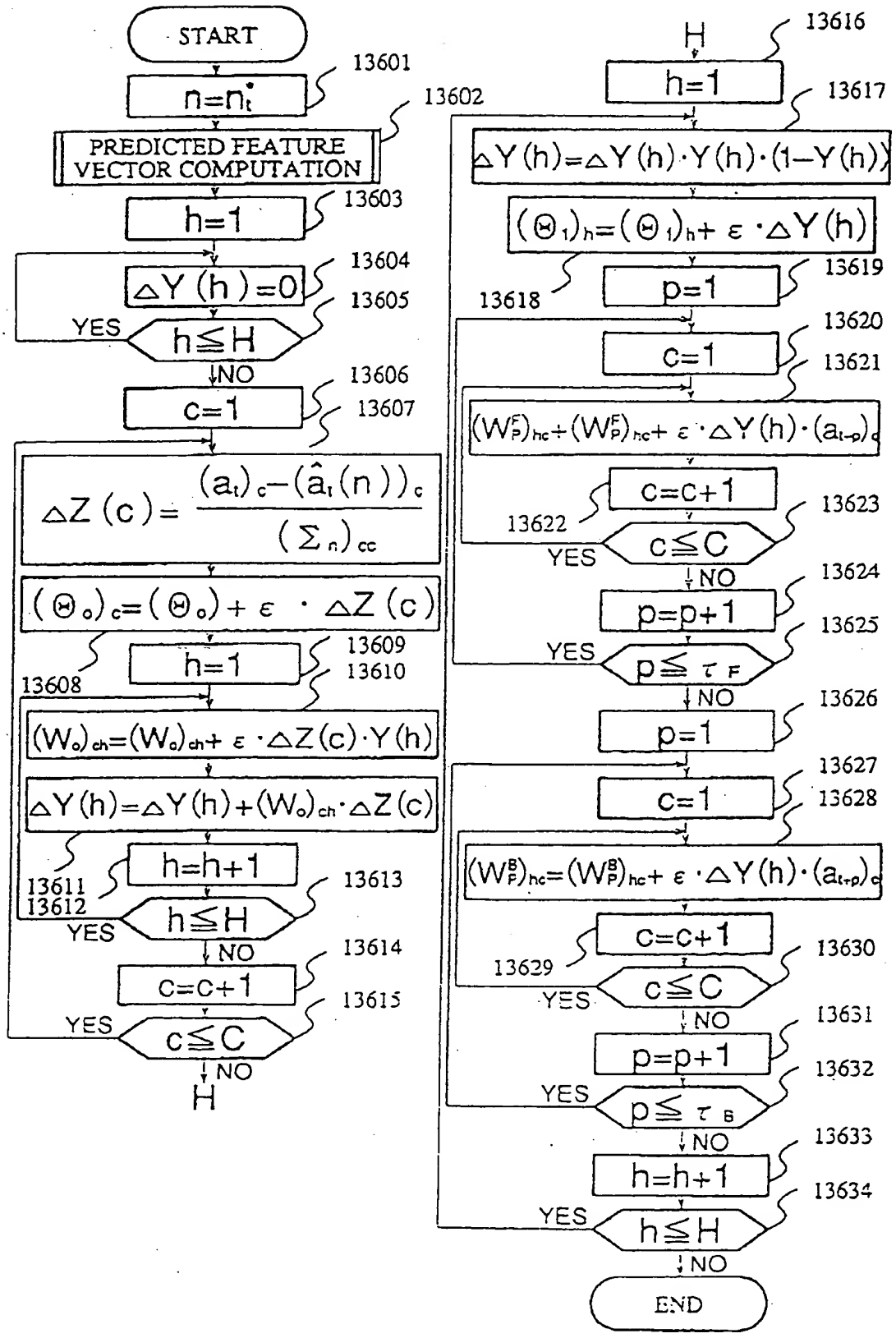

START
13601
$n=n_t^*$
13602
PREDICTED FEATURE VECTOR COMPUTATION
13603
h=1
13604
$\Delta Y(h)=0$
13605
YES
$h \leqq H$
NO
13606
c=1
13607
$\Delta Z(c)=\frac{(a_t)_c-(\hat{a}_t(n))_c}{(\Sigma_n)_{cc}}$
$(\Theta_o)_c=(\Theta_o)+\varepsilon \cdot \Delta Z(c)$
13608
13609
h=1
13610
$(W_o)_{ch}=(W_o)_{ch}+\varepsilon \cdot \Delta Z(c) \cdot Y(h)$
$\Delta Y(h)=\Delta Y(h)+(W_o)_{ch} \cdot \Delta Z(c)$
13611
13612
h=h+1
13613
YES
$h \leqq H$
NO
13614
c=c+1
13615
YES
$c \leqq C$
NO
H
H
13616
h=1
13617
$\Delta Y(h)=\Delta Y(h) \cdot Y(h) \cdot (1-Y(h))$
$(\Theta_1)_h=(\Theta_1)_h+\varepsilon \cdot \Delta Y(h)$
13618
13619
p=1
13620
c=1
13621
$(W_P^F)_{hc}=(W_P^F)_{hc}+\varepsilon \cdot \Delta Y(h) \cdot (a_{t-p})_c$
c=c+1
13622
13623
YES
$c \leqq C$
NO
13624
p=p+1
13625
YES
$p \leqq \tau_F$
NO
13626
p=1
13627
c=1
13628
$(W_P^B)_{hc}=(W_P^B)_{hc}+\varepsilon \cdot \Delta Y(h) \cdot (a_{t+p})_c$
c=c+1
13629
13630
YES
$c \leqq C$
NO
13631
p=p+1
13632
YES
$p \leqq \tau_B$
NO
13633
h=h+1
13634
YES
$h \leqq H$
NO
END

Fig. 14

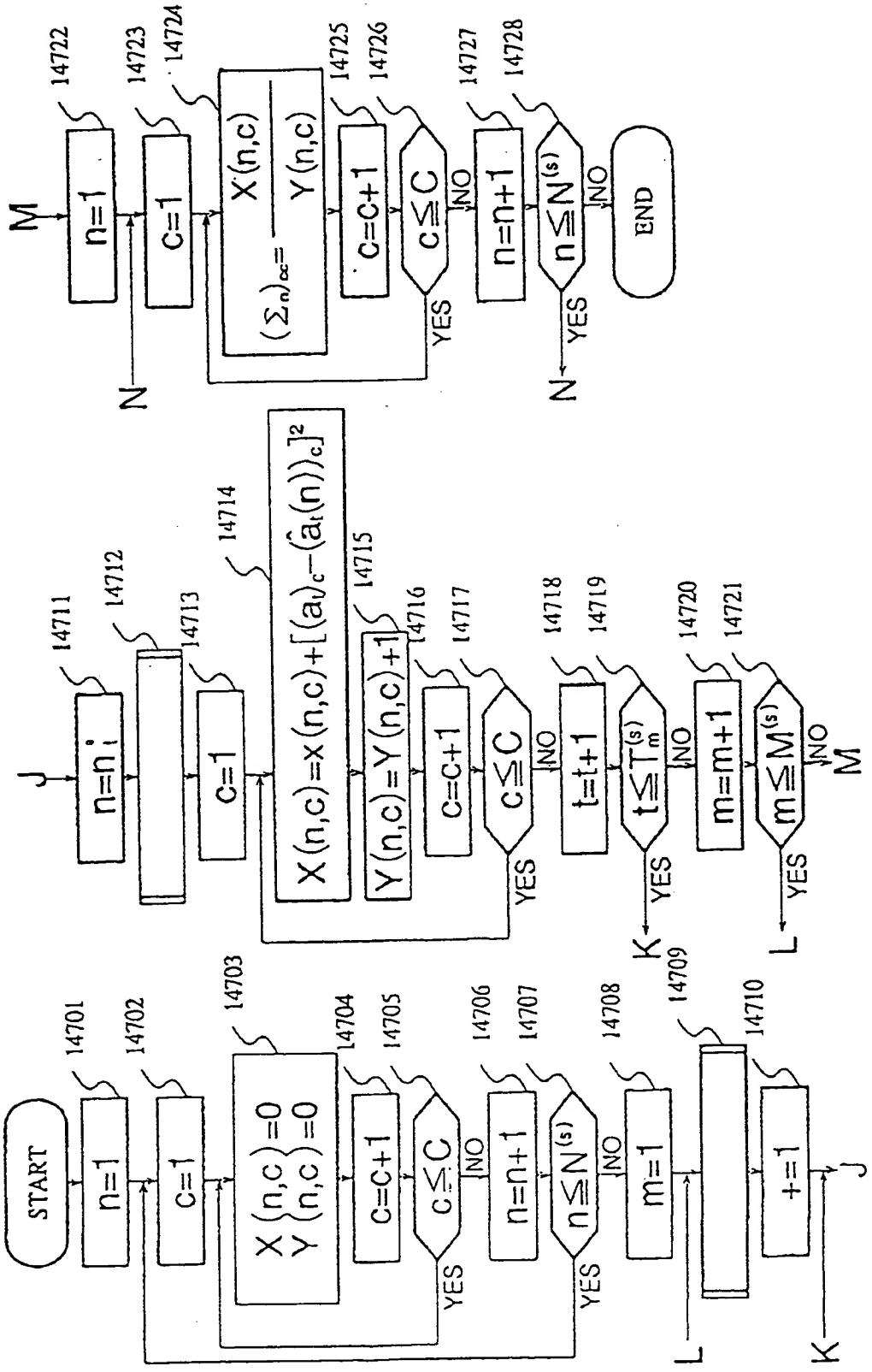
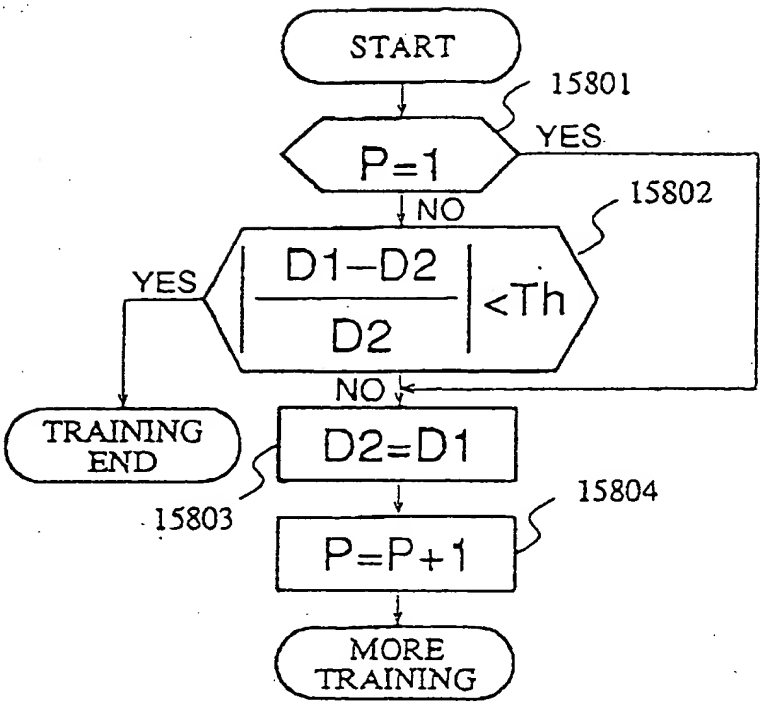

START
n=1
c=1
X(n,c)=0
Y(n,c)=0
c=c+1
c≦C
YES
NO
n=n+1
n≦N(s)
m=1
+=1
J
L
K
n=ni
X(n,c)=x(n,c)+[(a)c−(ât(n))c]²
Y(n,c)=Y(n,c)+1
t=t+1
t≦Tm(s)
m=m+1
m≦M(s)
M
N
X(n,c)/Y(n,c)
(Σn)αc=
END

Fig. 15


START
15801
P=1
YES
NO
15802
D1−D2 / D2 <Th
YES
NO
TRAINING END
D2=D1
15803
15804
P=P+1
MORE TRAINING

THIS PAGE BLANK (USPTO)

This Page is Inserted by IFW Indexing and Scanning Operations and is not part of the Official Record

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

- ☐ BLACK BORDERS
- ☐ IMAGE CUT OFF AT TOP, BOTTOM OR SIDES
- ☐ FADED TEXT OR DRAWING
- ☐ BLURRED OR ILLEGIBLE TEXT OR DRAWING
- ☐ SKEWED/SLANTED IMAGES
- ☐ COLOR OR BLACK AND WHITE PHOTOGRAPHS
- ☐ GRAY SCALE DOCUMENTS
- ☐ LINES OR MARKS ON ORIGINAL DOCUMENT
- ☐ REFERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY
- ☐ OTHER: _____

IMAGES ARE BEST AVAILABLE COPY.
As rescanning these documents will not correct the image problems checked, please do not report these problems to the IFW Image Problem Mailbox.

THIS PAGE BLANK (USPTO)